A Rough Set Approach to Feature Selection Based on Wasp Swarm Optimization

Huilian FAN 1,*, Yuanchang ZHONG 2

1 College of Mathematics and Computer Science, Yangtze Normal University, Fuling 408100, China
2 CCTTC in College of Communication Engineering, Chongqing University, Chongqing 400044, China

Abstract

As an important concept of rough set theory, an attribute reduction is a subset of attributes that are jointly sufficient and individually necessary for preserving a particular property of the given information table. In order to acquire minimal attribute reduction, we propose a wasp swarm optimization algorithm for attribute reduction based on rough set and the significance of feature. The significance of feature is constructed based on the mutual information between selected conditional attributes and decisional attributes. The algorithm dynamically calculates heuristic information based on the significance of feature to guide search. Experimental are carried out on some standard UCI datasets. The results demonstrate that, in terms of solution quality and computational effort, proposed algorithm can get better results than other intelligent swarm algorithms for attribute reduction.

Keywords: Rough Set; Attribute Reduction; Wasp Swarm Optimization; Information Entropy

1 Introduction

Attribute reduction is an important problem in the emerging field of data mining which is aimed at finding a small set of rules from the data set with predetermined targets. Rough set theory is one of the effective methods to attribute selection, which can preserve the meaning of the attributes[1]. The essence of rough set approaching to attribute reduction is the process of choosing a subset of attributes from the original set of features forming patterns in a given dataset. The subset should be necessary and sufficient to describe target concepts, retaining a suitably high accuracy in representing the original attributes. Unlike other dimensionality-reduction methods, minimal attribute reduction preserves the original meaning of the attributes after reduction[2].

In recent years, a lot of attribute reduction methods have been proposed. There are three key issues based attribute reduction algorithms: discernibility matrix[3], granular computing[4] and biology inspired algorithms[5, 6]. A discernibility-based method has been introduced to deal with attributes that are redundant or worse. Granular computing is a new intelligent computing...
theory and method based on issue partition. Swarm-based approaches such as ant colony and particle swarm optimization have been proved to be competitive in rough set attribute reduction areas. However, there are some shortcomings in these algorithms such as low speed and high space complexity, or leading to non-minimal feature combination, or premature convergence, etc.

To overcome these, we propose a new attribute reduction algorithm based on rough sets and WSO (Wasp Swarm Optimization) called ARWSO, which adopts mutual information and the significance of feature based information entropy of features as heuristic information for WSO.

This paper covers the implementation and the evaluation of the proposed algorithm. The rest of this paper is organized as follows. Sections 2, provides some preliminaries in rough theories relevant to feature selection and WSO. Section 3 details the proposed algorithm and the pseudo-code of our algorithm. The primary motivation of this paper is to discuss several benchmark problems from UCI and compare with other intelligent swarm algorithms for attribute reduction performance. Experimental results are given in Sections 4. Finally, the conclusions of this paper are described in Section 5.

2 Preliminaries

2.1 Rough set theory

Rough set theory\cite{7, 8, 9}, proposed by Pawlak is an approach to aid decision making in the presence of uncertainty. Here, we use the concept of rough set theory in term of data containing in an information table. The notion of information table provides a convenient tool for the representation of objects in terms of their attribute values.

An information table is a 4-tuple (quadruple) $S = \{U, C \cup D, V, f\}$, where $U$ is a non-empty finite set of objects, $A = C \cup D$ is a non-empty finite set of attributes, $V$ is the union of feature domains such that $V = \bigcup_{a \in A} V_a$, $V_a$ is the domain (value set) of attribute $a$, any $(u, a) \in U \times A$ determines a function $f(u, a) \in V_a$, called information (knowledge) function\cite{10, 11}.

**Definition 1** Let $S = \{U, C \cup D, V, f\}$ be an information table and $B \subseteq C \cup D$. Two elements $x_i, x_j \in U$ is said to be $B$-indiscernible if and only if $f(x_i, a) = f(x_j, a)$ for every $a \in B$, denoted by $\text{IND}(B)$.

\[
\text{IND}(B) = \{(x_i, x_j) \in U \times U | \forall a \in B, f(x_i, a) = f(x_j, a)\} \tag{1}
\]

The partition of $U$, generated by $\text{IND}(B)$ is denoted $U/\text{IND}(B)$ and can be calculated as follows:

\[
U/\text{IND}(B) = \bigotimes\{U/\text{IND}(\{a\})| a \in B\} \tag{2}
\]

where $\bigotimes$ is specifically defined as follows for sets $R$ and $S$:

\[
R \bigotimes S = \{X \cap Y| X \in R, Y \in S, X \cap Y \neq \emptyset\} \tag{3}
\]

2.2 Mutual information based feature selection

In feature selection problems, the relevant conditional features contain important information about decision feature, whereas the irrelevant features contain little information regarding decision. The task for feature selection is to find those conditional features that contain as much infor-
mation about decision as possible. For this purpose, Shannon’s information theory provides us a feasible way to measure the information of data set with entropy and mutual information[12, 13].

Let $S = \langle U, C \cup D, V, f \rangle$ be an information table. For any subset $R \subseteq C$ of features, $U/IND(R) = \{X_1, X_2, \cdots, X_n\}$ and $U/IND(D) = \{Y_1, Y_2, \cdots, Y_j\}$ denote the partitions induced by equivalence relations $IND(R)$ and $IND(D)$[6].

**Definition 2** The information entropy $H(R)$ of feature set $R$ is defined as

$$H(R) = -\sum_{i=1}^{n} p(X_i) \log p(X_i)$$

where $p(X_i) = |X_i|/|U|, 1 \leq i \leq n, |X_i|$ is the cardinality of $X_i$.

**Definition 3** The conditional entropy of $D$ conditioned to $R$ is defined as

$$H(D|R) = -\sum_{i=1}^{n} \left[ p(X_i) \sum_{j=1}^{m} p(Y_j|X_i) \log p(Y_j|X_i) \right]$$

where $p(X_i) = |X_i|/|U|, p(Y_j|X_i) = |X_i \cap Y_j|/|X_i|, 1 \leq i \leq n, 1 \leq j \leq m$.

**Definition 4** The mutual information between $R$ and $D$ is defined as

$$I(R; D) = H(D) - H(D|R)$$

If the mutual information is larger, the two feature sets are more closely related.

**Definition 5** Let $S = \langle U, C \cup D, V, f \rangle$ be an information table. For every $a \in C$, if $I(C - \{a\}; D) < I(C; D)$, then $a$ is a core feature of $S$.

**Definition 6** Let $S = \langle U, C \cup D, V, f \rangle$ be an information table. For any subset $R \subseteq C$ of features, if $I(R; D) = I(C; D)$ and for every $b \in R$, $I(R - \{b\}; D) < I(R; D)$, then $R$ is called a feature reduction of $C$ with respect to $D$ in information table $S$.

The problem of computing a minimal attribute reduction of a information table is converted into a nonlinear optimization problem meeting the following conditions[14].

$$\begin{cases} 
R \subseteq C \\
I(R; D) = I(C; D) \\
\forall a \in R, I(R - \{a\}; D) < I(R; D) \\
\text{min cardinality}(R)
\end{cases}$$

**Definition 7** Let $S = \langle U, C \cup D, V, f \rangle$ be an information table. For any subset $R \subseteq C$ of features, and any feature $a \in C - R$, the significance of feature $a$ with respect to $R$ and $D$ is defined as

$$sgn(a, R, D) = I(R \cup \{a\}) - I(R; D)$$

**2.3 Wasp swarm algorithm**

For years, researchers have been working to explain how simple insects like wasps can coordinate their behavior in order to build such complex nest structure. The foraging model of Wasp[15, 16]
describes the nature of interactions between an individual wasp and its local environment with respect to task allocation. They model the colony’s self-organized allocation of tasks with what they refer to as response thresholds. An individual wasp has a response threshold for each zone of the nest. Based on a wasp’s threshold for a given zone and the amount of stimulus from brood located, a wasp may or may not become engaged in the task of foraging for that zone.

Assume that \( m \) tasks need to be performed. These tasks are associated with stimuli or demands, the levels of which increase if they are not satisfied (because the tasks are not performed by enough individuals or at high enough rates)[17]. Let us assume that there are \( N \) wasps, denoted by \( i \), with response thresholds \( \theta_{ij} \) (\( i = 1, \ldots, N \) and \( j = 1, \ldots, m \)) with respect to task \( j \)-associated stimuli. Let \( s_j \) denote the intensity of task \( j \)-associated stimuli. In the fixed-threshold model, individual \( i \) engages in task \( j \) with probability

\[
T_{\theta ij}(s_j) = \frac{s_j^2}{s_j^2 + \theta_{ij}^2}
\]

For \( s_j \ll \theta_{ij}, T_{\theta ij}(s_j) \) is close to 0, and \( s_j \gg \theta_{ij}, T_{\theta ij}(s_j) \) is close to 1; at \( s_j = \theta_{ij}, T_{\theta ij}(s_j) = 0.5 \).

3 New Method for Feature Selection

3.1 Core features selection

The elements of feature core are those features that cannot be eliminated. The SelectFeatureCore algorithm for finding feature core is as follows.

Algorithm name: SelectFeatureCore

Input: an inform table \( S = \langle U, C \cup D, V, f \rangle \)

Output: core features \( Core \)

(1) Initial \( Core = \phi ; \)

(2) for every \( r \in C \)

(3) \( \{ \) if \( I(C - \{r\}; D) < I(C; D) \) then \( Core = Core \cup \{r\}; \) \}

(4)Output \( Core \)

An example dataset[18] is given in Table 1 to illustrate the core features selection algorithm. Here, the table consists of four conditional features \( (a, b, c, d) \), one decision feature \( (e) \), and eight objects.

From Table 1, based on each feature, there are five partitions of \( U \) induced by indiscernibility relation on each attribute, i.e. \( U/IND\{a\} = \{\{0, 3, 4\}, \{1, 7\}, \{2, 5, 6\}\}, U/IND\{b\} = \{\{0, 2, 4\}, \{1, 3, 6, 7\}, \{5\}\}, U/IND\{c\} = \{\{0, 4\}, \{1, 6, 7\}, \{2, 3, 5\}\}, U/IND\{d\} = \{\{0, 3\}, \{1, 2, 5, 6\}, \{4, 7\}\}, U/IND\{e\} = \{\{0\}, \{1, 3, 6\}, \{2, 4, 5, 7\}\}.

Based on formula in Eq.(4,5,6) and \( C = \{a, b, c, d\}, D = \{e\} \), the mutual information between \( C \) and \( D \) can be calculated as follows.

\[
I(C; D) = H(D) - H(D|C) = 0.423140 - 0 = 0.423140
\]

Using the same way, we obtain
Table 1: Example information table

<table>
<thead>
<tr>
<th>$x \in U$</th>
<th>$a$</th>
<th>$b$</th>
<th>$c$</th>
<th>$d$</th>
<th>$e$</th>
<th>$x \in U$</th>
<th>$a$</th>
<th>$b$</th>
<th>$c$</th>
<th>$d$</th>
<th>$e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>S</td>
<td>R</td>
<td>T</td>
<td>T</td>
<td>R</td>
<td>4</td>
<td>S</td>
<td>R</td>
<td>T</td>
<td>R</td>
<td>S</td>
</tr>
<tr>
<td>1</td>
<td>R</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>T</td>
<td>5</td>
<td>T</td>
<td>T</td>
<td>R</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>2</td>
<td>T</td>
<td>R</td>
<td>R</td>
<td>S</td>
<td>S</td>
<td>6</td>
<td>T</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>T</td>
</tr>
<tr>
<td>3</td>
<td>S</td>
<td>S</td>
<td>R</td>
<td>T</td>
<td>T</td>
<td>7</td>
<td>R</td>
<td>S</td>
<td>S</td>
<td>R</td>
<td>S</td>
</tr>
</tbody>
</table>

$I(C - \{a\}; D) = H(D) - H(D|C - \{a\}) = 0.423140$

$I(C - \{b\}; D) = H(D) - H(D|C - \{b\}) = 0.423140$

$I(C - \{c\}; D) = H(D) - H(D|C - \{c\}) = 0.423140$

$I(C - \{d\}; D) = H(D) - H(D|C - \{d\}) = 0.272625$

Because only $I(C - \{d\}; D) < I(C; D)$, we can obtain feature core of the given information table is $\{d\}$.

### 3.2 Heuristic information based attribute reduct

This paper adopts the idea of wasp swarm optimization and uses each foraging bee to search a possibility solution of minimal attribute reduction. Selecting each feature in our information table has an associated task. First, we get the feature core with mutual information, and then foraging bees search minimum attribute reduction on the current feature core by selecting the rest of features. The profitability selecting is a function of significance of the selecting feature based on current state.

Let $S = \langle U, C \cup D, V, f \rangle$, $R_i \subset C$, individual $i$ select feature $a \in \{C - R_i\}$ with probability:

$$P(a, R_i, D) = \frac{I(R_i \cup \{a\}; D) - I(R_i; D)}{\max_{j=1}^{[C-R_i]} (I(R_i \cup \{a_j\}; D) - I(R_i; D)) + \varepsilon}. \quad (9)$$

where $\varepsilon$ is a coefficient which determine the relative importance of the heuristic information. If $\varepsilon$ is more large, the wasps will make decision mainly based on randomness, and if $\varepsilon$ more small, respectively $\varepsilon = 0$, wasps will select those features with higher heuristic information in a greedy manner.

A search process is terminated by one of the following two conditions: (1) $I(R; D) = I(C; D)$, where $R$ is the current attribute reduction constructed by a wasp. (2) The cardinality of the current solution is larger than that of the temporary minimal attribute reduction.

The algorithm for minimal attribute reduction with wasp swarm and the significance of feature is as follows.

**Algorithm: AttributeReduce**

**Input:** information table $S = \langle U, C \cup D, V, f \rangle$

**Output:** a minimal feature reduct $R_{min}$ and its cardinality $L_{min}$
(1) initial $R_{min} = C$, $L_{min} = |C|$, $wasp\_numbers = |C|/6$;
(2) Core= \emptyset;
(3) for every $r \in C$
(4) \{ if ($I(C - \{ r \}; D) < I(C; D)$) then Core = Core $\cup \{ r \}$; \}
(5) for every $k \in wasp\_numbers$ { 
(6) $R_k = Core$, $L_k = |Core|$ 
(7) do{
(8) for every $a_k \in \{ C - R_k \}$
(9) \{ Compute probability $P(a_k, R_k, D)$ by Eq.(9); \} 
(10) Select next feature $b_k \in \{ C - R_k \}$ by roulette;
(11) $R_k = R_k \cup b_k$, $L_k = L_k + 1$;
(12) }until ($I(R_k; D) == I(C; D)$ or $L_k \geq L_{min}$) 
(13) if ($I(R_k; D) == I(C; D)$ and $L_k < L_{min}$)
(14) then $R_{min} = R_k$, $L_{min} = L_k$; 
(15) } Output $R_{min}$ and $L_{min}$

4 Experimental Results

In order to evaluate our algorithm called ARWSO, we compare the performance with IDSRSFS[5] and RSFSACO[6] algorithm, based on nine datasets which are from UCI datasets. The performance can be realized through the following goals: (1) The size of reduction of a data set after executing the AttributeReduct algorithm. (2) The run time from program start to termination.

Both IDSRSFS and RSFSACO parameter configuration is set according the requirements of reference [5, 6]. The number of wasp is one-sixth of condition attribute numbers and should not be less than 5.

We implement our algorithm for feature selection in Java. The computer is Intel Pentium 2.93 GHz CPU, 2G RAM and the operating system is Windows XP Professional. We tested the three algorithms on 9 UCI datasets.

The experimental results are summarized in Table 2. The leftmost column consists of dataset names. The 2\textsuperscript{nd} and 3\textsuperscript{rd} columns are instance numbers and feature numbers of corresponding data set. The 4\textsuperscript{th} column is the size of optimal reduct up to now. The 5\textsuperscript{th}, 6\textsuperscript{th}, 7\textsuperscript{th} and 8\textsuperscript{th} columns are results of IDSRSFS and RSFSACO algorithms. The rightmost column is result of our algorithm. These results are the size of reduction of test dataset of each algorithm finds and its average time. The unit of average time is ms. Each dataset is test for 20 times and the number in parentheses denotes the times of tests to achieve such a attribute reduct.

The results, core attributes and minimum attribute reduction, founded by our algorithm are given in Table 3. Each number in Table 2 denotes corresponding one column (one feature) of dataset. There are datasets that have more than one reduction result. For example, there are several minimum attribute reduction results for Lung dataset, such as $\{2 3 12 15\}, \{3 6 20 53\}$, etc. Table 3 only present one of these minimum reduction result.
From Table 2, we can see that ARWSO provides the best results, the performance of RSFSACO is better than IDSRSFS. We also found that all algorithms have similar efficiency when they handle datasets with less than 16 features. However, while the number of attribute increased, the probably of finding minimum attribute reduction by IDSRSFS become smaller, but RSFSACO and ARWSO also can find out them in great probability. This is likely due to IDSRSFS’s drawback - not having heuristic information to search through the feature space for optimal solutions and premature convergence to a local optimum in the space. Due to ARWSO and RSFSACO search the optimum solution base on core, and less computing for foraging behavior of wasp colony than ant colony, average run time of ARWSO is less than RSFSACO, IDSRSFS.

Table 2: Experimental results comparison with IDSRSFS and RSFSACO

<table>
<thead>
<tr>
<th>Data</th>
<th>Inst</th>
<th>Feat</th>
<th>Min</th>
<th>IDSRSFS</th>
<th>Avetime</th>
<th>IDSRSFS</th>
<th>Avetime</th>
<th>IDSRSFS</th>
<th>Avetime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monk1</td>
<td>124</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>478</td>
<td>3</td>
<td>310</td>
<td>3</td>
<td>302</td>
</tr>
<tr>
<td>Tic-tac-toe</td>
<td>958</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>3315</td>
<td>8</td>
<td>2917</td>
<td>8</td>
<td>2821</td>
</tr>
<tr>
<td>Zoo</td>
<td>101</td>
<td>16</td>
<td>5</td>
<td>5</td>
<td>997</td>
<td>5</td>
<td>616</td>
<td>5</td>
<td>580</td>
</tr>
<tr>
<td>Vote</td>
<td>435</td>
<td>16</td>
<td>9</td>
<td>9</td>
<td>10634</td>
<td>9</td>
<td>9283</td>
<td>9</td>
<td>8699</td>
</tr>
<tr>
<td>Lung</td>
<td>32</td>
<td>56</td>
<td>4</td>
<td>4</td>
<td>17282</td>
<td>4</td>
<td>13027</td>
<td>4</td>
<td>11983</td>
</tr>
<tr>
<td>Soybean_small</td>
<td>47</td>
<td>35</td>
<td>2</td>
<td>2</td>
<td>3941</td>
<td>2</td>
<td>3529</td>
<td>2</td>
<td>3216</td>
</tr>
<tr>
<td>Dermatology</td>
<td>366</td>
<td>34</td>
<td>6</td>
<td>8</td>
<td>75166</td>
<td>6</td>
<td>56477</td>
<td>6</td>
<td>53132</td>
</tr>
<tr>
<td>kr-vs-kp</td>
<td>3196</td>
<td>36</td>
<td>29</td>
<td>31</td>
<td>951935</td>
<td>29</td>
<td>852762</td>
<td>29</td>
<td>800287</td>
</tr>
<tr>
<td>Audiology</td>
<td>200</td>
<td>69</td>
<td>13</td>
<td>16</td>
<td>261029</td>
<td>13</td>
<td>218332</td>
<td>13</td>
<td>193796</td>
</tr>
</tbody>
</table>

Table 3: The result of core attributes and minimum attribute reduction using our algorithm

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Core attributes</th>
<th>Minimum attribute reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monk1</td>
<td>{1 2 5}</td>
<td>{1 2 5}</td>
</tr>
<tr>
<td>Tic-tac-toe</td>
<td>φ</td>
<td>{1 2 3 4 5 6 7 9}</td>
</tr>
<tr>
<td>Zoo</td>
<td>{6 13}</td>
<td>{4 6 8 12 13}</td>
</tr>
<tr>
<td>Vote</td>
<td>{1 2 3 9 11 13 16}</td>
<td>{1 2 3 4 9 11 13 15 16}</td>
</tr>
<tr>
<td>Lung</td>
<td>φ</td>
<td>{3 12 13 14}</td>
</tr>
<tr>
<td>Soybean_small</td>
<td>φ</td>
<td>{22 23}</td>
</tr>
<tr>
<td>Dermatology</td>
<td>φ</td>
<td>{3 4 16 20 32 34}</td>
</tr>
<tr>
<td>kr-vs-kp</td>
<td>{1 3 4 5 6 7 10 12 13 15 16 17 18 20 21 23 24 25 26 27 28 30 31 33 34 35 36}</td>
<td>{1 3 4 5 6 7 10 11 12 13 15 16 17 18 20 21 22 23 24 25 26 27 28 30 31 33 34 35 36}</td>
</tr>
<tr>
<td>Audiology</td>
<td>{1 6 47}</td>
<td>{1 2 4 5 6 7 10 11 14 15 47 64 66}</td>
</tr>
</tbody>
</table>
5 Experimental Results Analysis

This paper proposes a new algorithm based on rough sets and WSO in order to overcome drawbacks of failing to find optimal reductions. Our algorithm utilizes mutual information based information entropy to found core attributes, and then utilizes the significance of feature as probability information to search through the feature space for minimum attributes reduction result. Our algorithm has the following characteristics: (a) It applies the foraging behavior of wasp swarm to feature selection; (b) It constructs probability information based on the basis of mutual information and the significance of feature.

Acknowledgement

The authors gratefully acknowledge the reviewer for their valuable comments, and the support of the Natural Science Project of Chongqing Education Commission under Grant No. KJ111304.

References


